Video Generation

Naeemullah Khan
naeemullah.khan@kaust.edu.sa



جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

KAUST Academy
King Abdullah University of Science and Technology

August 9, 2026

Table of Contents

1. Learning outcomes
2. Introduction: video generation today and the path from media synthesis to world simulators
3. World models: motivation, definition, and the classic V–M–C architecture (Ha & Schmidhuber)
4. Generative vs. joint-embedding paradigms and how world models are used
5. Generative world models: **Sora** (history, DiT architecture, limitations)
6. Text-driven video editing: **Text2LIVE** (layered neural atlases, CLIP supervision)
7. Interactive generative video: **Genie-1** (tokenizer, latent actions, dynamics, robotics)
8. **Genie-2** and **Genie-3**: scale, capabilities, and limitations
9. Latent predictive models: **V-JEPA** and **V-JEPA 2**
10. Lecture summary and references

⇒ Learning Outcomes

By the end of this session, you should be able to:

- State what a world model is, $s' \sim p(s' \mid s, a)$, and explain why prediction rather than search is what makes fast control possible.
- Describe the classic Vision–Memory–Controller architecture and how each module is trained.
- Contrast pixel-generative stacks with latent-predictive models, and say which one a given task calls for.
- Explain how an interactive video model can obtain an action space from video that carries no action labels.
- Identify the limitations that current systems share: rollout duration, closed action sets, and physical plausibility.

Video Generation Today: Stunning Visuals



Image from OpenAI's Sora website.

Video Generation: From media synthesis to world simulators

- A useful split today: systems that prioritize high-fidelity pixels and instruction following (e.g. large text-to-video diffusion) versus systems that prioritize compact latent dynamics for control (e.g. JEPA-style prediction).
- Video remains expensive: long tensors and coherent motion are harder than sharp stills; evaluation must separate perceptual quality from physical plausibility.
- The frontier is shifting from realistic-looking clips to models that approximate causal structure and action consequences, what curricula often bundle under *world models*.

Four paradigms (curriculum-style taxonomy)

Paradigm	Architectural idea	Representative systems	Primary objective
Structured 3D scenes	NeRF / Gaussian splatting; differentiable rendering	DreamFusion, GSGen, Marble (World Labs)	Multi-view–consistent 3D that routes through graphics pipelines
Interactive generative video	Action-conditioned diffusion or autoregressive transformers	Genie, GameNGen, GAIA (Wayve)	Interactive "game-like" rollouts from actions
Latent predictive	Joint embedding predictive architectures (JEPA); latent dynamics, with a decoder only if pixels are wanted	V-JEPA, PAN	Predict future *representations*; V-JEPA never decodes pixels, PAN adds a video diffusion decoder
Text-to-video simulators	Large DiT / diffusion stacks at scale	Sora, Veo, Gen-3	Maximize fidelity + prompt alignment; emergence as "simulator"

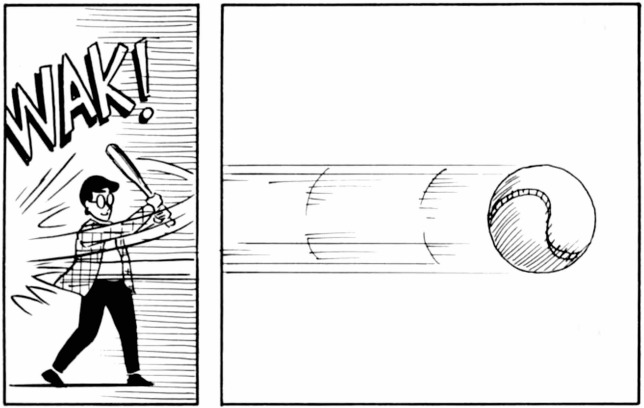
Pedagogical splits echo materials such as CMU 10-423/10-723 generative-AI world-model lectures and CVPR-style world-model tutorials; see references.

World Models

How generative AI can model, imagine, and predict entire environments and their dynamics.

World Models: Motivation

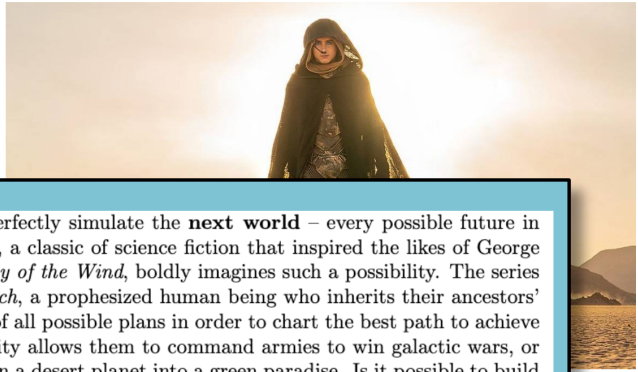
Intuition: a baseball batter has milliseconds to decide how to swing, shorter than the time visual signals take to reach the brain. Action must be driven by a **predictive**, action-conditioned internal model rather than by slow deliberate search.



Scott McCloud excerpt (time–space in comics).

World Models: Motivation

Dune offers one notion of a world model...



What would you do if you could perfectly simulate the **next world** – every possible future in the environment that we reside? *Dune*, a classic of science fiction that inspired the likes of George Lucas' *Star Wars* and Miyazaki's *Valley of the Wind*, boldly imagines such a possibility. The series is centered around the *Kwisatz Haderach*, a prophesized human being who inherits their ancestors' memories and simulates the outcomes of all possible plans in order to chart the best path to achieve their goals [25]. Such superhuman ability allows them to command armies to win galactic wars, or to oversee global-scale projects that turn a desert planet into a green paradise. Is it possible to build towards computer systems with similar functionalities using a similar approach?

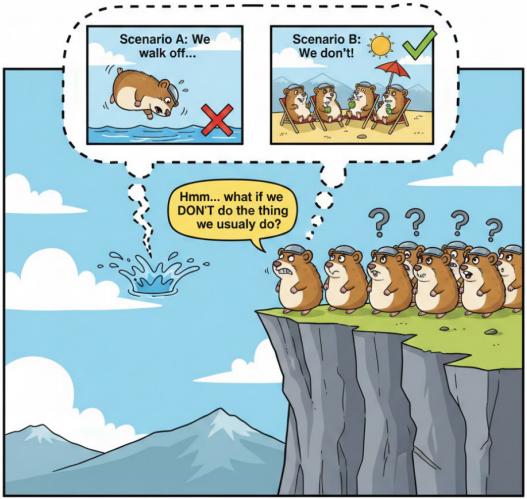
Text from http://arxiv.org/abs/2507.05169

CMU 10-423, Lecture 26.

World Models: Motivation

the psychology literature has long defined our notion of a world model as “hypothetical thinking”

informally, we speak about this as a “thought experiment”



CMU 10-423, Lecture 26.

World Models: Motivation

- Humans develop a mental model of the world based on what they are able to perceive with their limited senses. Our decisions and actions are based on this internal model.
- Train a generative world model in an unsupervised way to learn a compact spatio-temporal representation of the environment.
- Use features from the world model to train a simple policy (agent) to solve the required task.

World Models: Definition

Definition: *A world model* takes a previous world state s and an action a, and samples (or predicts) the next world state s' through a distribution (or function):

$$s' \sim p(s' \mid s, a)$$

World Models: Definition

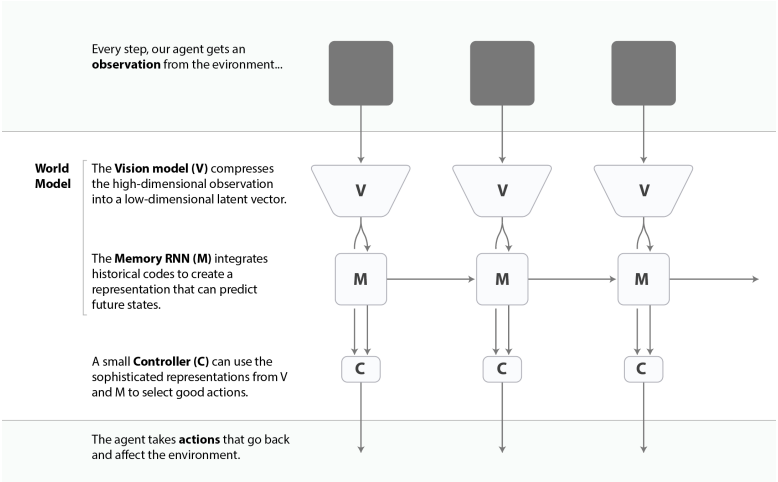
Question: What does a world state hold?

Answer: It depends on what you're trying to model. Consider the contrast of:

A) All geopolitical entities, their leaders, and decisions they make

B) Physical interactions in a 3D world

World Models: Overview

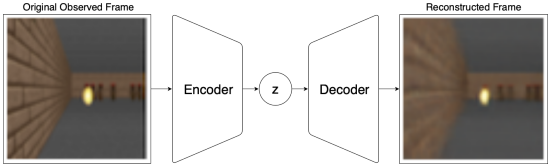


World model overview.

Source: Ha & Schmidhuber (2018); arXiv:1803.10122.

Vision module: variational autoencoder (V)

- **Goal:** learn an abstract, compressed representation of each observed input frame.
- **VAE:** first stage of a two-stage sequence model, similar in spirit to modern tokenizers but trained with the classic VAE objective.

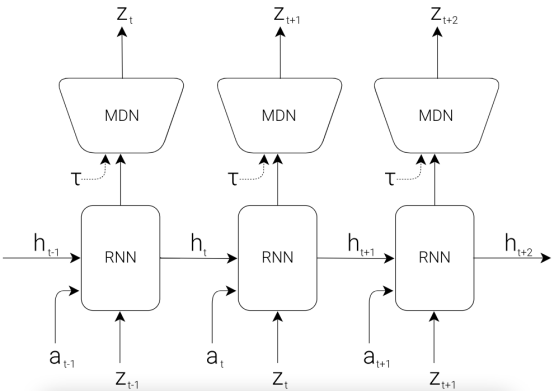


VAE diagram.

Source: Ha & Schmidhuber (2018); arXiv:1803.10122.

Memory module: Predict the future

- Model $p(z_{t+1} \mid z_t, a_t, h_t)$ where z_t is the latent state at time t, a_t is the action at time t, and h_t is the hidden state at time t.
- adjust a temperature parameter τ to control model uncertainty



MDN-RNN architecture.

Source: Ha & Schmidhuber (2018); arXiv:1803.10122.

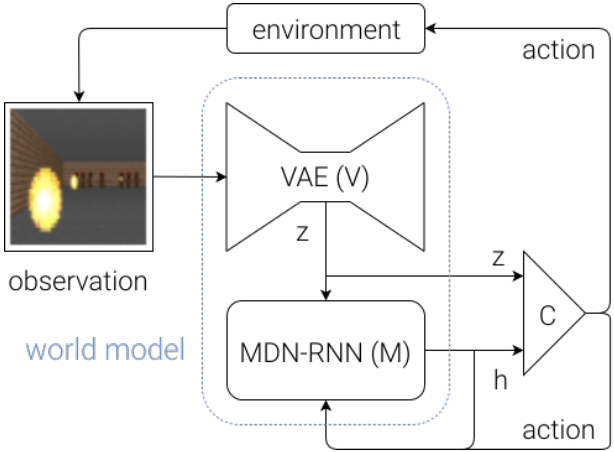
Controller module: decision-making (C)

- The **Controller (C)** selects actions to maximize expected cumulative reward during agent rollout.
- Designed to be as simple and small as possible; trained **separately** from V and M. Because C holds so few parameters it is **evolved with CMA-ES** (Covariance-Matrix Adaptation Evolution Strategy) on cumulative reward, not backpropagated end-to-end.
- Most agent complexity is thus concentrated in the world model (V and M).
- C is a **single-layer linear** model that maps the latent vector z_t and RNN hidden state h_t directly to the action a_t at each time step:

$$a_t = W_c[z_t; h_t] + b_c$$

- Here, W_c and b_c are the weight matrix and bias; $[z_t; h_t]$ is the concatenated latent and memory state.
- Output vector a_t is the action taken by the agent.

World model: Putting Everything Together



Agent-environment flow (observation $\rightarrow z_t$, RNN state h_t, action a_t, next step).

Source: Ha & Schmidhuber (2018); arXiv:1803.10122.

World Model Architectures: Generative vs. Joint Embedding

- **Generative: predicts y** (with all the details, including irrelevant ones)
- **Joint Embedding: predicts an abstract representation of y**

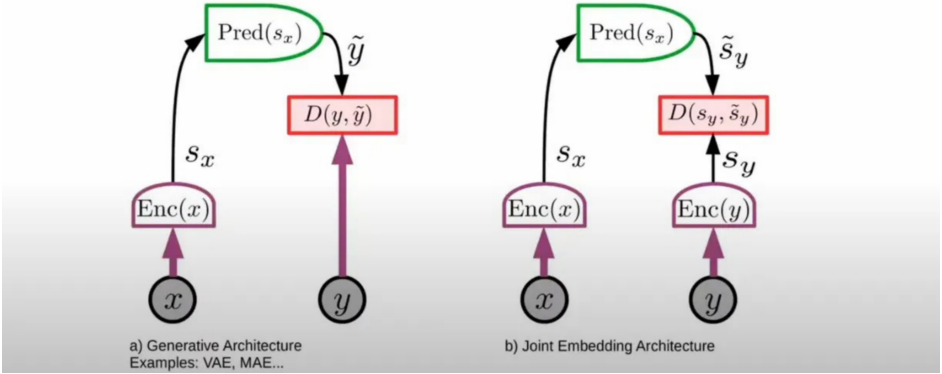


Image credit: CS25.

How Could We Use a World Model?

Generate a 3D Scene, then Render/Simulate





Interactive Videos





Latent World Representations





CMU 10-423, Lecture 26.

Generative World Models

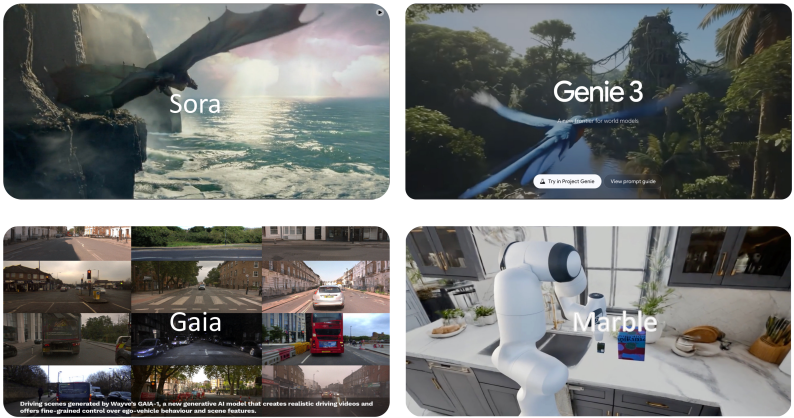


Image credit: CS25.

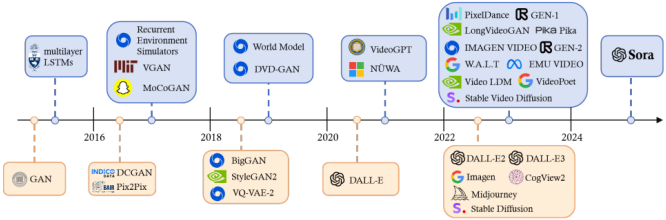
Generative World Models: Sora

- Generates high-quality, minute-long videos with strong visual coherence across many kinds of physical and digital scenes.
- Rapidly turns text prompts into compelling videos, accelerating creative work, education, and accessibility.
- Enables new applications in marketing, game development, and more through scalable video generation.



Liu et al. (2024); arXiv:2402.17177

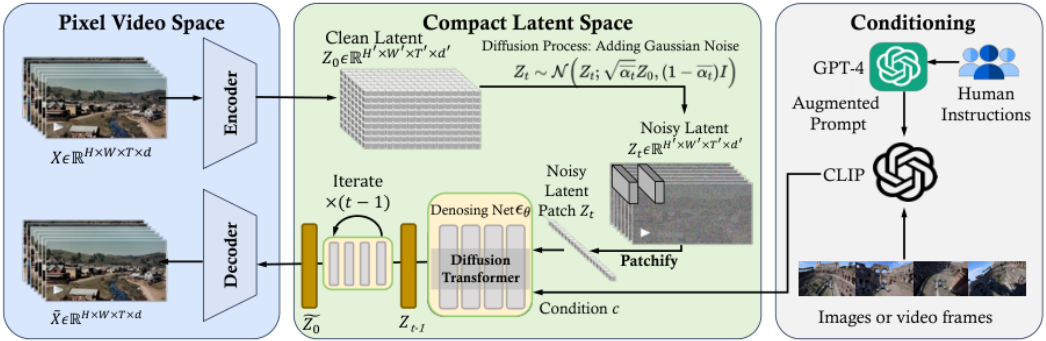
A Brief History: From Texture Synthesis to Sora



Liu et al. (2024), arXiv:2402.17177, Fig. 3.

- **Before deep learning:** Image generation relies on hand-crafted features.
- **The deep learning revolution:** GANs and VAEs brought greater realism to images; diffusion models later became the backbone of modern visual generation.
- **Transformers and modality fusion:** Transformers, adapted from NLP (BERT, GPT) to vision tasks (e.g. ViT), enable new capabilities; multimodal models (e.g. CLIP) align vision and language for T2I and T2V generation.
- **Sora and beyond:** Models like Sora represent the frontier, generating high-quality, minute-long videos from text prompts.

Sora Core: Diffusion Transformer Architecture



Liu et al. (2024), arXiv:2402.17177, Fig. 4.

- A **diffusion transformer (DiT)** replaces the U-Net denoiser of earlier diffusion models with a transformer over latent patches (Peebles & Xie, ICCV 2023), which is what makes the stack scale.
- Time-space compressor: compresses input video into compact latent space representation.

Sora Core: Diffusion Transformer Architecture (cont.)

- The transformer processes tokenized spacetime latents and predicts denoised versions.
- CLIP-like conditioning: incorporates user instructions and visual prompts, guiding generation with learned theme and style.
- Denoising and decoding: after multiple denoising steps, final latent video decoded back into pixel space to yield generated video.

Sora: Limitations

- **Physical realism:** unstable causal detail (e.g. objects not updating consistently after contact); implausible rigid-body motion; occasional absurd multi-object interactions.
- **Spatial/temporal control:** confusion over placement and left/right; drift in camera motion and event order; stray characters or animals that break the intended shot.
- **Interaction & prompting:** fine-grained local edits remain difficult; complex natural-language constraints may be misread.
- **Access:** closed red-team access at survey time (Feb 2024); Sora launched publicly in Dec 2024, with Sora 2 in 2025.

Text2LIVE

Text2LIVE (Bar-Tal et al., ECCV 2022) edits *existing* images and videos from a text prompt: instead of generating a video from scratch, it changes the appearance of objects or adds semi-transparent effects while keeping the rest of the footage intact.

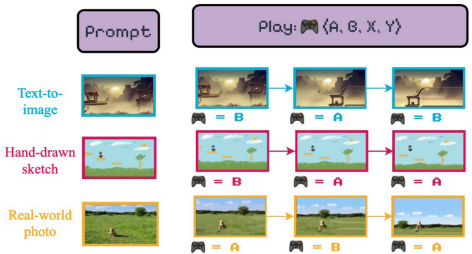
- **Layered editing:** the edit is generated on a separate RGBA layer that is composited over the original frames, so the source content is preserved.
- **Neural atlases:** the video is decomposed into 2D atlases (foreground/background); editing the atlas propagates the change consistently across all frames.
- **CLIP supervision, zero-shot:** the edit layer is optimized so the result matches the text prompt under CLIP similarity; no paired training data; instead a generator is trained per input (internal learning) at test time.
- **Contrast with generation:** complements the text-to-video simulators seen earlier; here language steers *modification* of real footage rather than synthesis from noise.

Text2LIVE (cont.)



[Text2LIVE: Text-Driven Layered Image and Video Editing, Bar-Tal, Ofri-Amar, Fridman, Kasten & Dekel, ECCV 2022]

Genie-1: interactive video generation



Source: CMU Lecture 26 World Models

Genie-1 is an interactive video generation model.

At test time:

- Takes a prompt (image) as input
- At each time step, accepts an action given by the user
- Given the current frame & action, generates the next frame

This simulates the style of interaction one would have with a video game.

2D Platformer Games (Setting)

Q: What is a 2D Platformer game?

A: A 2D platformer is a video game where the player moves a character through a flat, two-dimensional world, usually by running, jumping, and avoiding obstacles or enemies. The gameplay often focuses on timing, movement precision, and reaching goals such as exits, items, or new areas.

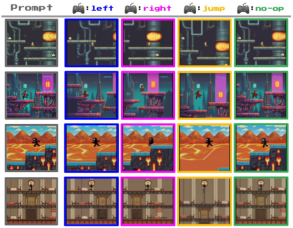
Canonical Example:
Super Mario World for SNES



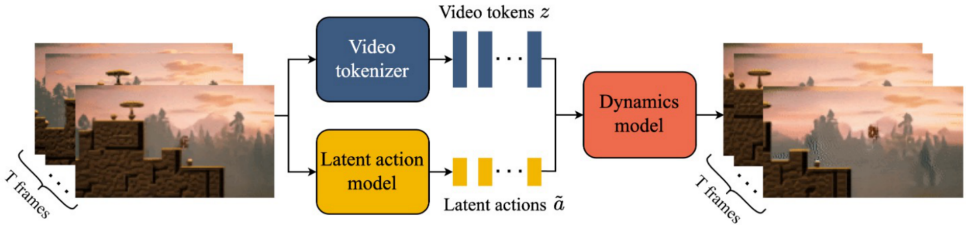
Space Panic: The first 2D platformer.

Source: CMU 10-423, Lecture 26 (World Models); screenshot of *Space Panic* (Universal, 1980).

Genie-1: training data



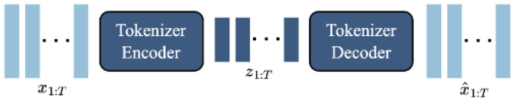
- **Primary training data:** 6.8M 16s videos (30k hours) of 2D platformer games
- **Model:** three components
  1. Video tokenizer/detokenizer
  2. Latent action model
  3. Dynamics model
- **Size:** 11B parameters



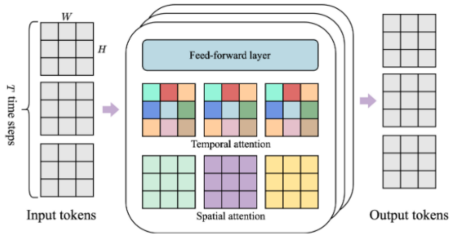
Source: CMU Lecture 26, World Models.

Genie-1: video tokenizer

Video tokenizer / detokenizer:



ST-Transformer:



Video tokenizer / detokenizer:

- VQ-VAE that encodes *T* frames of video:

$$\boldsymbol{x}_{1:T} = (x_1, x_2, \cdots, x_T) \in \mathbb{R}^{T \times H \times W \times C}$$

- into *T* discrete tokens (from vocab of size D)

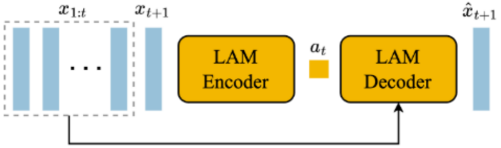
$$\boldsymbol{z}_{1:T} = (z_1, z_2, \cdots, z_T) \in \mathbb{I}^{T \times D}$$

- Trained via standard VQ-VAE objective over video dataset
- *Backbone model:*
  - ST-Transformer, which scales well for longer sequences

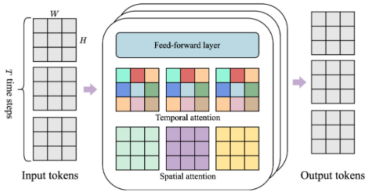
Source: CMU Lecture 26, World Models.

Genie-1: latent action model

Latent Action Model:



ST-Transformer:



Latent Action Model (LAM):

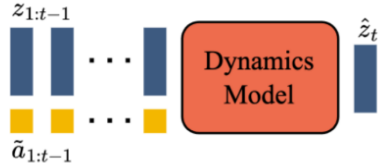
- An unsupervised model for learning *latent* actions from unlabled videos
- Another VQ-VAE model
  - Encoder: converts t frames of video plus the t+1 frame ($\mathbf{x}_{1:t+1}$) into sequence of t+1 actions ($\mathbf{a}_{1:t+1}$)
  - Decoder: converts t frames ($\mathbf{x}_{1:t}$) and t actions ($\mathbf{a}_{1:t}$) into the t+1 frame ($\mathbf{x}_{t+1}$)
- At test time, this model is discarded since actions come from a human user

- *Backbone model:* ST-Transformer

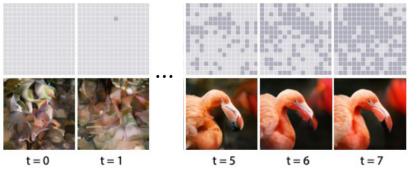
Source: CMU Lecture 26, World Models.

Genie-1: dynamics model (MaskGIT)

Latent Action Model:



Decoder-only MaskGIT:



Dynamics Model

- Decoder-only MaskGIT
- Conditions on the sequence of previous video tokens ($z_{1:t}$) and actions ($\mathbf{a}_{1:t}$) to generate the next video token (z_{t+1}), which can be detokenized into a video frame ($\mathbf{x}_{t+1}$)
- Trained via cross-entropy for next token prediction
- *Note*: Action embeddings are incorporated **additively** in the dynamics model (and LAM), rather than via **concatenation**
- *Backbone model:* ST-Transformer

Source: CMU Lecture 26, World Models.

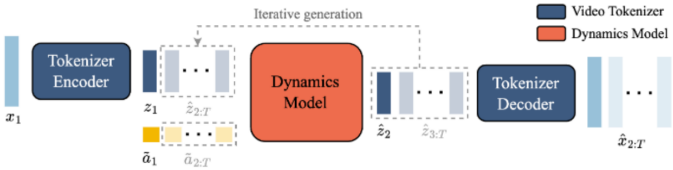


Figure 8 | **Genie Inference**: the prompt frame is tokenized, combined with the latent action taken by the user, and passed to the dynamics model for iterative generation. The predicted frame tokens are then decoded back to image space via the tokenizer's decoder.

Source: CMU Lecture 26, World Models.

Genie-1 Qualitative Results

- Images created with a text-to-image model can serve as a prompt.
- Then Genie-1 enables a user to interact directly with the environment.
- The latent actions are mapped to actual keys (a, s, d, f, etc.) so that the user can explore what they do.

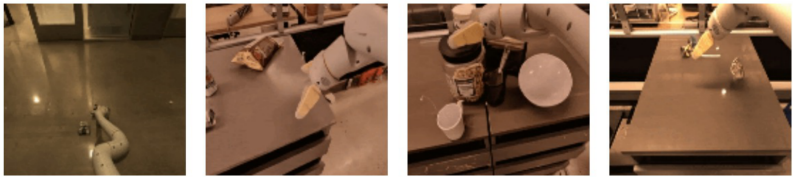


Source: CMU Lecture 26, World Models.

Genie-1 Application to Robotics

- **Robotics dataset:**
  - 130k robotics demonstrations
  - a simulation dataset (size unspecified)
  - 209k episodes of real robot data
- No actions included, just videos

- Training Genie on the robotics datasets learns a new model that can simulate "game" style play with a robotic arm as shown below



latent actions: 0, 0, 1, 2, 6, 1, 1, 4, 0, 0

Source: CMU Lecture 26, World Models.

Genie-1: Are the Latent Actions Transferable?

- CoinRun is a procedurally generated platformer.
- **Models:**
  1. **Behavior Cloning:** train a policy given a dataset of each frame paired with the expert action (skyline)
  2. **Random Agent:** always take a random action (baseline)
  3. **LAM-based Policy:** train a policy given a dataset of each frame paired with the Genie latent action; separately learn a mapping from latent actions to expert actions from a few examples

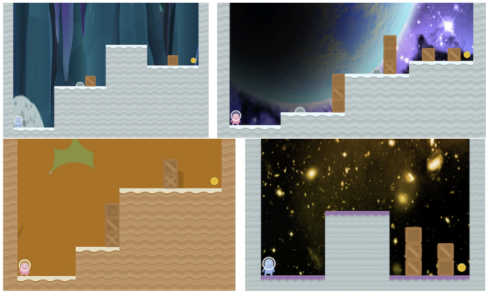


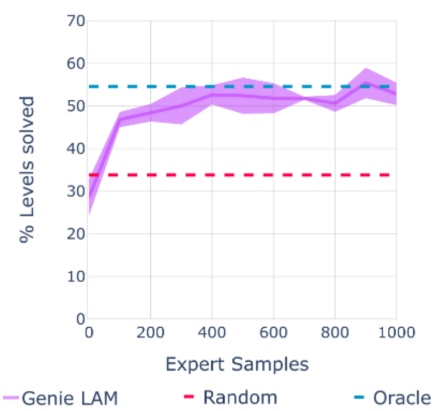
Figure: Four generated CoinRun levels with the coin at the right.

Source: CMU 10-423, Lecture 26 (World Models);

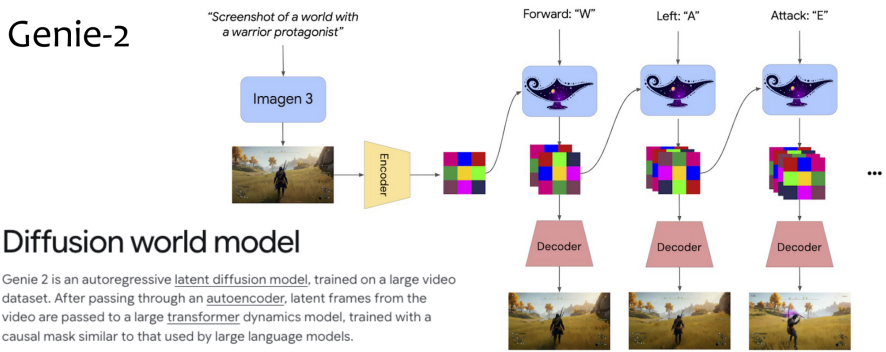
CoinRun environment from Cobbe et al., arXiv:1812.02341.

Genie-1: Are the Latent Actions Transferable? (Result)

- **Yes.** The LAM-based policy reaches the same score as the expert-action skyline (plotted as *Oracle*) given as few as **200** action-labelled samples to fit the latent-to-real mapping, in both the easy and hard settings.
- Both stay far above the random agent.
- Metric: mean percentage of levels solved out of 100 samples, averaged over 5 seeds.
- So latent actions learned from unlabelled Internet video carry over to an environment the model never trained on, which is what makes them useful for control.



Source: CMU Lecture 26, World Models; result per Bruce et al. (arXiv:2402.15391), Fig. 15 (Easy).

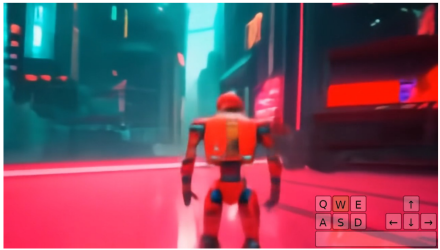


Source: CMU Lecture 26, World Models.

Genie-2

- The technical description of Genie-2 is rather vague.
- This could indicate that the primary change is **scale**:
  - **more parameters**
  - **more training data**





Source: CMU Lecture 26, World Models.

Genie-3

	GameNGen	Genie 2	Veo	Genie 3
Resolution	320p	360p	720p to 4K	720p
Domain	Game-specific	3D Environments	General	General
Control	Game-specific	Limited keyboard / mouse actions	Video-level description*	Navigation; Promptable world events
Interaction Horizon	A few seconds	10-20 seconds	8 seconds	Multiple minutes
Interaction Latency	Real time	Not real time	N/A	Real time

*Additional controls such as references, style and camera are available.

Source: CMU Lecture 26, World Models.



Source: CMU Lecture 26, World Models.

Genie-3: Limitations

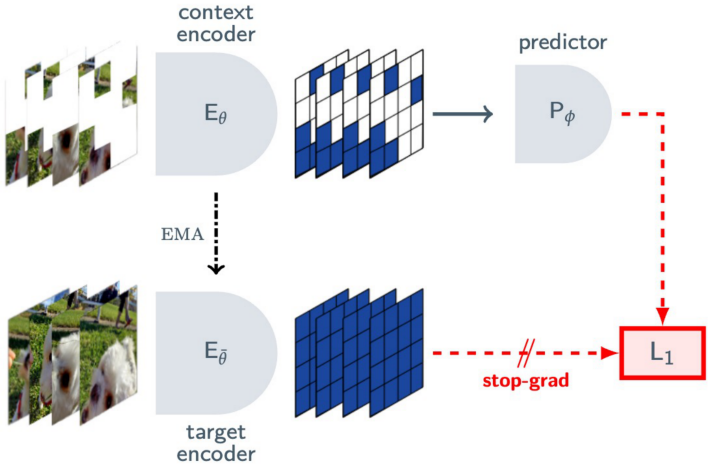
- Interaction window is limited to a short duration (a few minutes).
- Actions are (likely) latent and from a fixed set; defining new actions is not possible.
- Prompting for changes in the world is possible and represents an open set; this creates an asymmetry between the closed action set and the open world change set.
- Limited to one character (at least that seems to be the case).
- Fundamentally just a video model: no game-engine-interpretable world representation (e.g., point cloud, mesh, Gaussian splatting) is provided.
- Extremely high computational demands.

V-JEPA: predicting representations instead of pixels

- **JEPA** = *Joint Embedding Predictive Architecture*. Mask part of a video in space and time, then predict the **representation** of the masked region from the visible context.
- The loss lives in representation space, so there is no decoder and no pixel reconstruction. V-JEPA is trained **solely** on feature prediction: no pretrained image encoder, no text, no negative pairs, no reconstruction term.
- **Why bother:** pixels contain detail that is genuinely unpredictable (exact texture, noise). A reconstruction loss spends capacity on it; a representation-space loss can ignore it.
- Trained on 2M public videos, largest model ViT-H/16. Frozen-backbone evaluation: 81.9% Kinetics-400, 72.2% Something-Something-v2, 77.9% ImageNet-1K.
- **The catch:** nothing here is action-conditioned, so V-JEPA predicts *what comes next*, not *what happens if I act*. That is what V-JEPA 2 adds.

Bardes et al., arXiv:2404.08471.

V-JEPA: video JEPA



Bardes et al., arXiv:2404.08471. Image credit: CS25.

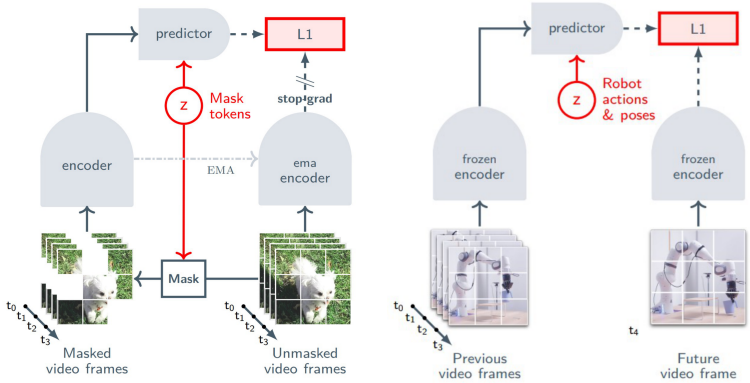
V-JEPA 2: from prediction to planning

- **V-JEPA 2-AC** is the action-conditioned variant: the predictor is additionally given the action, so it models the consequence of acting rather than just the continuation of a clip.
- It is post-trained on under **62 hours** of unlabelled robot video from the Droid dataset, which is small next to the 2M-video pretraining corpus.
- Given an image of the goal, the model is used for **planning**: search over candidate actions, roll them forward in latent space, and pick the one whose predicted representation lands nearest the goal.
- Deployed **zero-shot** on Franka arms for reaching, grasping and pick-and-place in environments not seen during training, with no task-specific reward or demonstration.
- This is the payoff of the latent route: planning needs a state you can compare against a goal, and it never needed the pixels.

Assran et al., *V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning*; arXiv:2506.09985.

V-JEPA 2: action-conditioned latent prediction

Action Conditioned → World Model



Assran et al., arXiv:2506.09985. Image credit: CS25.

Lecture summary

- **Motivation:** video is the natural domain for dynamics and "world" hypotheses; we contrast **pixel generative** stacks with **latent predictive** models.
- **World Models:** Can be used to generate interactive videos, 3D scenes and latent world representations.
- **The Simulation Gap:** Despite stunning visuals (e.g., pouring liquid), current models lack symbolic or true physics-based understanding (e.g., accurate fluid dynamics).
- **Takeaway:** The two families answer different questions. Ask a pixel model for a convincing clip; ask a latent model for a state you can plan against. Which one you need is set by the task, not by sample quality.
- **Looking ahead:** The open problems are duration and controllability: rollouts stay coherent for minutes rather than hours, and action sets remain closed while the space of requested world changes is open.

References

1. From Video Generation to World Model (CVPR 2025 tutorial). world-model-tutorial.github.io
2. Survey: Is Sora a World Simulator? (arXiv:2405.03520). link
3. Gormley & Nayebi (CMU 10-423 / 10-723, *Generative AI*): Interactive World Models, Lecture 26 slides. PDF
4. Ha: World Models interactive write-up / demos. worldmodels.github.io
5. Bardes et al. (V-JEPA). arXiv:2404.08471
6. Assran et al. (V-JEPA 2). arXiv:2506.09985
7. Meta AI: Introducing V-JEPA 2. link
8. Bruce et al. (Genie). arXiv:2402.15391
9. Sora review paper (arXiv:2402.17177). link

References (cont.)

10. Bar-Tal, Ofri-Amar, Fridman, Kasten & Dekel (Text2LIVE), ECCV 2022. arXiv:2204.02491
11. Peebles & Xie: Scalable Diffusion Models with Transformers (DiT), ICCV 2023. arXiv:2212.09748
12. Bar-Tal et al. (Lumiere). arXiv:2401.12945; project site
13. Learning World Models for Interactive Video Generation (arXiv:2505.21996). link
14. GameNGen project. site; arXiv:2408.14837
15. Hafner et al.: Dream to Control (Dreamer; arXiv:1912.01603). link
16. Hafner et al.: DreamerV3 (arXiv:2301.04104). link
17. Ha & Schmidhuber: World Models (arXiv:1803.10122). link
18. Cobbe, Klimov, Hesse, Kim & Schulman: Quantifying Generalization in Reinforcement Learning (CoinRun). arXiv:1812.02341
19. Stanford CS25 deck used on slides (Google Drive). link